THE LANCET Infectious Diseases

Supplementary webappendix

This webappendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Jing Q-L, Liu M-J, Zhang Z-B, et al.Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort. *Lancet Infect Dis* 2020; published online June 17. https://doi.org/10.1016/S1473-3099(20)30471-0.

Web Appendix

1.1 Case definition, case identification and contact tracing

All the investigations Guangzhou CDC conducted on COVID-19 patients, asymptomatic infections of SARS-CoV-2 and their close contacts are in accordance with the Prevention and Control Plan for the Novel Coronavirus Pneumonia (editions 1-7) issued by the National Health Commission of China. The case definitions are similar across the 2nd -5th editions issued on Jan.22 – Feb. 21, which covers most of our study period (the first patients was identified by Guangzhou CDC on Jan. 21, and the last symptom onset in our data was on Feb. 14, 2020). The only differences are 1) reporting of asymptomatic infections was required since the 3rd edition (issued on Jan. 28); and 2) "fever" was changed to "fever and/or respiratory symptoms" as one of the clinical criteria for suspected cases since the 4th edition (issued on Feb. 7). The following description is based on the 5th edition.

Case definition A suspected case is defined as patients meeting ≥1 epidemiological criteria and ≥2 clinical criteria outlined below.

Epidemiological criteria

(1) Travel or residence history in Wuhan or nearby cities during 14 days before symptom onset;
(2) Contact history with a PCR-positive COVID-19 patient during 14 days before symptom onset;
(3) Contact history during 14 days before symptom onset with patients who had fever or respiratory symptoms and came from Wuhan or communities with reported COVID-19 cases; and
(4) Related to a cluster of COVID-19 cases.

Clinical criteria

(1) Fever and/or respiratory symptoms
(2) Radiographic characteristics of pneumonia, such as multiple ground-glass shadows, infiltrative shadows and consolidation in both lungs;

(3) normal or lower leukocyte counts, or lower lymphocyte counts at acute phased of the disease.

A confirmed case is defined as a suspected patient with positive detection of 2019-nCoV nucleic acid by real-time RT-PCR or viral genes that are highly homologous to 2019-nCoV by sequencing using respiratory specimens. An asymptomatic infection is an individual with laboratory confirmation but without clinical signs. Asymptomatic infections were reported as "test-positive" rather than "confirmed" in the national surveillance system but the status would be changed to "confirmed" if symptoms developed and were reported.

An imported case is defined as a case who had residence in or travel history to Hubei Province (where Wuhan is located) during the 2 weeks before symptom onset; otherwise, this case is considered as a local case.

To define primary and secondary cases, we first identify the earliest symptom onset date in each case cluster and call that day 0. A local case with symptom onset on days 0 or 1 are considered as a local primary cases. An imported case with symptom onset on days 0-3 is considered as an imported primary case. A cluster may have multiple co-primary cases. This difference in time thresholds is based on the belief that imported cases from Hubei Province are more likely to be the source of infection of the cluster. Here we assume that an imported case with symptom onset more than 3 days later than the earliest symptom onset in the cluster is unlikely to be the source of infection, which is reasonable given that the probability of a serial interval < -3 days is very small.[1] All other cases are considered as either imported or local secondary cases, depending on whether they meet the definition of imported or local cases. Imported primary cases and local primary cases are both primary cases. Likewise, imported secondary cases and local secondary cases are both secondary cases. Asymptomatic infections are also classified as primary or secondary using the collection date of the first nasal swab that was tested positive. The four categories, imported primary, local primary, imported secondary and local secondary, are distinguished from each other in the calculation of effective reproductive numbers.

Case ascertainment, reporting and epidemiological investigation COVID-19 cases were ascertained mainly by two routes: (1) medical institutes are required to screen patients with fever, dry cough and short breath for unknown reasons and ask patients whether they have residence or travel history to Wuhan and surrounding cities or communities that have reported

cases, contact history with individuals with fever or respiratory symptoms from above areas, contact history with individuals infected with the novel coronavirus, or linked to clusters of COVID-19 patients, during the 14 days prior to their symptom onset; (2) community authorities are required to screen residents or visitors for residence or travel history to Wuhan and surrounding cities or communities that have reported cases during the past 14 days and for respiratory symptoms, fever, chill, fatigue, diarrhea, or conjunctival congestion. Community authorities are required to report to local public health authorities if such individuals are identified.

Medical institutes should report identified suspected cases, confirmed cases and asymptomatic infections to the internet-based national novel coronavirus reporting system if feasible or to the county CDC within two hours. Upon notification by medical institutes, county CDC should report initial data to the internet-based national novel coronavirus reporting system immediately. County CDC should finish epidemiological investigation within 24 hours and upload the case investigation form (attached to the end of this appendix) to the national reporting system in two hours after the completion of the investigation.

Medical institutes need to collect clinical specimens and send the specimens to designated local or provincial CDC labs or third-party laboratories for testing as soon as possible. Specimens to be collected may include upper respiratory samples (e.g., nasal swabs), lower respiratory samples (e.g., sputum from deep cough, Alveolar lavage fluid), fecal samples, anal swabs, and blood samples.

Contact tracing, monitoring and testing A close contact is defined as any individual who was within 1 meter from a suspected or confirmed case within 2 days before symptom onset or an asymptomatic infection within 2 days before specimen collection, if such contact is made in the absence of personal protective equipment. Specific types of close contact include: (1) Individuals living, learning, working or performing other types of activities in close distance to the COVID-19 patient, e.g., in the same classroom or household; (2) Healthcare providers or family members who treat, care or visit the COVID-19 patient, or patients and their visitors in the same ward; (3) Passengers and crew members who shared ride on any transportation vehicle with the COVID-19 patient; and (4) Other individuals who are considered by the investigators as close contacts.

Close contacts were quarantined and monitored at designated facilities such as hotels if feasible or at home otherwise for 14 days counting from the last unprotected contact with patients or asymptomatic infections. Nasal swabs were collected twice, one at day 1 and the other near day 14. The samples were tested by real-time RT-PCR at Guangzhou CDC or county-level CDC. Body temperature was measured twice each day, and a medical monitoring form (attached to the end of this appendix) was filled daily. If samples were tested positive or if any symptom (fever, chill, dry cough, wet cough, short breath, nasal congestion, runny nose, sore throat, fatigue, myalgia, headache, conjunctival congestion, joint pain, nausea, vomiting, diarrhea, etc.) was noticed, the individual was sent to a designated clinic for evaluation and sample collection. If diagnosed as a suspected or confirmed case or an asymptomatic infection, this individual was then managed and reported as described above for cases.

1.2 Calculating the effective reproductive number

Effective reproductive number, denoted as R_t, is a popular metric for assessing the temporal trend of the transmissibility of infectious diseases.[2,3] However, existing methods were not designed for cluster data.[4,5] In this analysis, we then use a simple moving average approach to estimate the effective reproductive number R_t for the period of Jan. 16-Feb. 6, 2020, assuming all secondary cases in a cluster were infected by the primary cases in that cluster. Calculation is limited to this period because primary or secondary cases outside this period are relatively few and uncertainty will be too large. For each day, the general estimator for R_t we use is given by:

$$\hat{R}_t = \frac{N_{sec}(t-2, t+2)}{N_{pri}(t-2, t+2) + \widetilde{N}_{sec}(t-2, t+2)},$$

where $N_{pri}(t_1, t_2)$ and $\widetilde{N}_{sec}(t_1, t_2)$ are the total numbers of primary and secondary cases in all clusters whose onset dates were within the time window $[t_1, t_2]$, and $N_{sec}(t_1, t_2)$ is the total numbers of secondary cases who might have been infected during the same window. The denominator and the numerator capture the numbers of potential infectors and infectees between whom the transmissions likely occurred during the interval $[t-2, t+2]$. A few secondary cases in the numerator could be tertiary cases and should probably be allocated to the next interval; however, by the same token, some secondary cases in the numerator of the previous interval should be allocated to the current interval. Therefore, whether tertiary cases are properly allocated does not affect much the estimation of R_t. The calculation of $N_{pri}(t_1, t_2)$, $N_{sec}(t_1, t_2)$

and $\widetilde{N}_{sec}(t_1, t_2)$ differs by how we allocate imported secondary cases and local primary cases to the numerator or the denominator, but there are two **common principle assumptions:** (1) **secondary cases in a cluster were infected by the primary cases in the same cluster; and (2) imported primary cases were infected outside guangzhou**. Analogous to $N_{pri}(t_1, t_2)$, $N_{sec}(t_1, t_2)$ and $\widetilde{N}_{sec}(t_1, t_2)$, we define $N_{pri}^{imp}(t_1, t_2)$, $N_{sec}^{imp}(t_1, t_2)$ and $\widetilde{N}_{sec}^{imp}(t_1, t_2)$ for imported primary and secondary cases, as well as $N_{pri}^{loc}(t_1, t_2)$, $N_{sec}^{loc}(t_1, t_2)$ and $\widetilde{N}_{sec}^{loc}(t_1, t_2)$ for local primary and secondary cases. Specifically, $N_{sec}^{imp}(t_1, t_2)$ and $N_{sec}^{loc}(t_1, t_2)$ are the numbers of imported and local secondary cases in the clusters whose primary cases are included in $N_{pri}^{imp}(t_1, t_2)$ and $N_{pri}^{loc}(t_1, t_2)$. Unlike $N_{pri}(t_1, t_2)$, $N_{sec}(t_1, t_2)$, and $\widetilde{N}_{sec}(t_1, t_2)$, these category-specific numbers are observed and fixed. R_t is calculated in the following three scenarios:

(1) All imported cases, regardless of primary or secondary, are considered as primary cases in the denominator, and secondary cases in the numerator only include local secondary cases.

$$N_{sec}(t-2, t+2) = N_{sec}^{loc}(t-2, t+2)$$
$$N_{pri}(t-2, t+2) = N_{pri}^{loc}(t-2, t+2) + N_{pri}^{imp}(t-2, t+2) + N_{sec}^{imp}(t-2, t+2)$$
$$\widetilde{N}_{sec}(t-2, t+2) = \widetilde{N}_{sec}^{loc}(t-2, t+2)$$

(2) Same as (1), but local primary cases on each day is allocated to previous days according to the assumed distribution of the incubation period and contribute to the numerator for those days. The rationale is that these local primary cases might have been infected by other cases in the previous days.

$$N_{sec}(t-2, t+2) = N_{sec}^{loc}(t-2, t+2) + \sum_{\tau=t-1}^{T} N_{pri}^{loc}(\tau, \tau) \sum_{s=t-2}^{t+2} \eta(\tau - s)$$
$$N_{pri}(t-2, t+2) = N_{pri}^{loc}(t-2, t+2) + N_{pri}^{imp}(t-2, t+2) + N_{sec}^{imp}(t-2, t+2)$$
$$\widetilde{N}_{sec}(t-2, t+2) = \widetilde{N}_{sec}^{loc}(t-2, t+2)$$

where $\eta(l)$ is the probability that the incubation period is l days (see section 1.4). We use the setting of $\eta(l)$ with a mean of 5 days given in Table S1.

(3) Same as (2), but imported secondary cases are now considered as secondary cases, not primary cases, and contribute to both $N_{sec}(t-2, t+2)$ in the numerator and $\widetilde{N}_{sec}(t-2, t+2)$ in the denominator.

$$N_{sec}(t-2,t+2) = N_{sec}^{loc}(t-2,t+2) + N_{sec}^{imp}(t-2,t+2) + \sum_{\tau=t-1}^{T} N_{pri}^{loc}(\tau,\tau) \sum_{s=t-2}^{t+2} \eta(\tau - s)$$

$$N_{pri}(t-2,t+2) = N_{pri}^{loc}(t-2,t+2) + N_{pri}^{imp}(t-2,t+2)$$

$$\widetilde{N}_{sec}(t-2,t+2) = \widetilde{N}_{sec}^{loc}(t-2,t+2) + \widetilde{N}_{sec}^{imp}(t-2,t+2)$$

The confidence interval (CI) for R_t is calculated based on the CI for the Poisson mean: $\exp\left[\log(\widehat{R_t}) \pm 1.96 \times \left(N_{pri}(t-2,t+2)\widehat{R_t}\right)^{-1/2}\right]$. We expect the magnitudes of the R_t for the three scenarios to be ordered as $(1) \leq (2) \leq (3)$. Scenarios 1 and 3 serve as the lower and upper bounds for the R_t. Local primary cases were likely infected locally and should be accounted for as infectors in the estimation of R_t (scenarios 2 and 3). Meanwhile, some imported secondary cases might be infected locally in Guangzhou (scenario 3). Consequently, the truth more likely lies between scenarios 2 and 3.

1.3 The chain-binomial model

The general model We use a discrete-time chain-binomial model to analyze the transmission process among close contacts.[6] Suppose we observe H transmission close contact groups (CCG), each with one or more primary cases with their household and/or non-household contacts, along with specific contact history between each pair of individuals. Let n_h be the size of CCG h, and let $N = \sum_{h=1}^{H} n_h$ be the total number of the people in these CCGs. We use Λ_h to represent the collection of individuals in CCG h. We consider two types of person-to-person close contact: frequent contact between household members and opportunistic contact between cases and non-household individuals. The probability that an infectious case infects a household member per daily contact is p_1, and the probability that an infectious case infects a non-household individual per daily period is p_2. In addition, each susceptible individual is subject to a constant daily infection probability of b via casual contact with the general public. Let $\tilde{t}_i$ be the symptom onset day if an infected person i is symptomatic, and we assume $\tilde{t}_i$ marks the peak infectivity during the infectious period. If person i is an asymptomatic infection, we also assign a $\tilde{t}_i$ which is interpreted as the peak infectivity day. Subjects who are not infected by their last observation day T_i will have $\tilde{t}_i = \infty$. Consider the potential transmission between an infectious person j and a susceptible person i in CCG h. Let $c_{ij}(t)$ indicate whether there is a household contact (1) or a non-household contact (2) or no contact at all (0) between individuals i and j. Let $I_{\{c\}}$ be the

indicator function that takes value 1 (0) if the condition c is true (false). The probability $p_{ji}(t)$ that a fully infectious individual j infects a susceptible individual i on day t is determined by

$$\text{logit}\left(p_{ij}(t)\right) = I_{\{c_{ij}(t)\neq 0\}}\left[I_{\{c_{ij}(t)=1\}}\text{logit}(p_1) + I_{\{c_{ij}(t)=2\}}\text{logit}(p_2) + \boldsymbol{\beta}'\boldsymbol{x}_{ij}(t)\right], \quad (1)$$

where the logit function has the form $\text{logit}(y) = \log\left(y/(1-y)\right)$, and $\boldsymbol{x}_{ij}(t)$ is the vector of covariates and $\boldsymbol{\beta}$ is the corresponding coefficient vector ($\boldsymbol{\beta}'$ means transpose of $\boldsymbol{\beta}$). In our notation, all bolded symbols are column vectors. Similarly, the covariate-adjusted infection probability by the casual contact with the general public is given by

$$\text{logit}\big(b_i(t)\big) = \text{logit}(b) + \boldsymbol{\alpha}'\boldsymbol{x}_i(t).$$

In practice, $\boldsymbol{x}_i(t)$ is usually a subset of $\boldsymbol{x}_{ij}(t)$, as the latter also encodes covariates associated with the transmission source j. For example, $\boldsymbol{x}_i(t)$ may be the age group and gender of individual i, which may also appear in $\boldsymbol{x}_{ij}(t)$. It is also common to assume coefficients in $\boldsymbol{\alpha}$ coincide with the corresponding coefficients in $\boldsymbol{\beta}$, as covariates of person i should modify his or her susceptibility in the same way regardless of the transmission source. The probability that a susceptible individual $i \in \Lambda_h$ escapes infection from all infective sources on day t is then given by $e_i(t) = \big(1 - b_i(t)\big)\sum_{j\in\Lambda_h}\phi(t-\tilde{t}_j)\theta^{1-s_j}p_{ij}(t)$, where $\phi(t-\tilde{t}_j)$ is the probability that individual j is infectious on day t given that j has symptom onset or peak infectivity on day $\tilde{t}_j$, which is also referred to as the relative infectivity function. Note that we assume that $\phi(l)$ depends on $l = t - \tilde{t}_j$, the distance between t and $\tilde{t}_j$, and that $D_{min} \le l \le D_{max}$, where D_{min} and D_{max} are the lower and upper bound of the infectious period in reference to the symptom onset or peak infectivity day. Note that $l = 0$ corresponds to the symptom onset day. D_{min} can be either positive, 0 or negative. If $\phi(l) > 0$ when $l < 0$, it implies infectivity during the incubation period. The symbol s_j indicates whether person j is a symptomatic case ($s_j = 1$) or an asymptomatic infection (s_j=0), and θ measures the relative infectivity level of an asymptomatic infection in comparison to a symptomatic case. θ can be estimated if there are a sufficient number of asymptomatic infections but is assumed known if not so. It is difficult, if not impossible, to estimate θ and $\phi(l)$ in the presence of other unknown parameters (p_1, p_2 and $\boldsymbol{\beta}$); consequently, we assume they are both known. We assume $\theta = 1$ and infectivity of an asymptomatic infection is the same as that during the incubation period of a symptomatic case,

and we perform sensitivity analysis by changing the values of $\phi(l)$. Denote the probability of individual i escaping infection up to day t by $Q_i(t) = \sum_{\tau=1}^{t} e_i(\tau)$. Suppose individual i is infected on day t. We assume the incubation period $d_i = \tilde{t}_i - t$ has a known discrete distribution $\eta(l) = Pr(d_i = l)$, $d_{min} \le l \le d_{max}$, where d_{min} and d_{max} are the minimum and maximum duration of the incubation period. If individual i is an asymptomatic infection, δ_i cannot be called incubation period, but we assume δ_i follows the same distribution. The distribution of the incubation period will be derived from literature. Define $\tilde{\boldsymbol{t}}_h = \{\tilde{t}_i : i \in \Lambda_h\}$ as the collection of symptom onset days (or peak infectivity days for asymptomatic infections). If $\tilde{\boldsymbol{t}}_h$ is fully observed, we can construct the likelihood contributed by person i as

$$L_{(i)}(b, p_1, p_2, \boldsymbol{\beta} | \tilde{\boldsymbol{t}}_h) = \begin{cases} Q_i(T_i), & \text{if not infected,} \\ \sum_{t=\tilde{t}_i - d_{min}}^{\tilde{t}_i - d_{max}} \{\eta(t - (\tilde{t}_i - d_{max}))[1 - e_i(t)]Q_i(t-1)\}, & \text{if infected.} \end{cases}$$

It is important to note that the design of our study is called case-ascertained design, which implies that each CCG is observed because of the primary case in the CCG. For this type of design, the infection of primary cases will not contribute to the likelihood, but their infectious periods contributed to the risk of infection of their group members and thus contributed to the likelihoods of those individuals. For simplicity, in each CCG, day 1 in the likelihood corresponds to the actual day $\tilde{t}_{\Lambda_h} - d_{max} + 1$, where $\tilde{t}_{\Lambda_h}$ is the earliest symptom onset day among primary cases of the CCG. For proper inference, one has to condition the likelihood on the fact that all group members who are not primary cases have not had symptom onset by day $\tilde{t}_{\Lambda_h}$. Under this condition, day $\tilde{t}_{\Lambda_h} - d_{max} + 1$ is the first day with uncertainty about the infection status of a group member who is not the primary case.[6]

The E-M algorithm For asymptomatic infections, we do not observe symptom onset. According to our assumption, there exists a peak infectivity day $\tilde{t}_i$ and that the time lag between $\tilde{t}_i$ and the infection time t follows the same distribution of the incubation period that is assumed known for symptomatic cases. For simplicity, we also call this time lag the incubation period. However, we do not observe $\tilde{t}_i$, and we use an Expectation-Maximization (EM) algorithm to integrate out such uncertainty.[7] Briefly, let $\boldsymbol{U}_h$ be the collection of $\tilde{t}_i$ for all individuals $i \in \Lambda_h$ and who are asymptomatic infections. In this analysis, we set the range of possible $\tilde{t}_i$ for each asymptomatic

infection to be from $\tilde{t}_{\Lambda_h} - 1$ to the last observation date of the study, Feb. 18, 2020, and will perform sensitivity analysis by varying this range. Let $\boldsymbol{O}_h$ be the collection of $\tilde{t}_i$ for all individuals $i \in \Lambda_h$ and who are symptomatic cases. $\boldsymbol{U}_h$ and $\boldsymbol{O}_h$ represent the unobserved and observed outcomes in group h. Let $\boldsymbol{u}_{hl}$, $l = 1, \cdots, \delta_h$ be all possible realizations of $\boldsymbol{U}_h$, where δ_h is the number of such realizations. Let $\boldsymbol{\psi} = (b, p_1, p_2, \boldsymbol{\beta})$ denote all the parameters. Rewrite the individual likelihood as $L_{(i)}(\boldsymbol{\psi}|\boldsymbol{O}_h, \boldsymbol{U}_h)$, and define the group-level and population-level likelihood based on complete data as $L_h(\boldsymbol{\psi}|\boldsymbol{O}_h, \boldsymbol{U}_h) = \prod_{i\in\Lambda_h} L_{(i)}(\boldsymbol{\psi}|\boldsymbol{O}_h, \boldsymbol{U}_h)$ and $L(\boldsymbol{\psi}|\boldsymbol{O}, \boldsymbol{U}) = \prod_{h=1}^{H} L_h(\boldsymbol{\psi}|\boldsymbol{O}_h, \boldsymbol{U}_h)$, respectively, where $\boldsymbol{O} = \{\boldsymbol{O}_h, h = 1, \cdots, H\}$ and $\boldsymbol{U} = \{\boldsymbol{U}_h, h = 1, \cdots, H\}$ represent population-level observed and missing outcomes. The EM algorithm proceeds as follows.

1. Choose an initial value $\boldsymbol{\psi}^{(0)}$, and set $\widehat{\boldsymbol{\psi}}^{(0)} = \boldsymbol{\psi}^{(0)}$.
2. At iteration $r \geq 1$, update the conditional probabilities $\lambda_{hk}^{(r)} = \frac{L_h(\widehat{\boldsymbol{\psi}}^{(r)}|\boldsymbol{O}_h, \boldsymbol{u}_{hk})}{\sum_{l=1}^{\delta_h} L_h(\widehat{\boldsymbol{\psi}}^{(r)}|\boldsymbol{O}_h, \boldsymbol{u}_{hl})}$, $k = 1, \cdots, \delta_h$. For CCGs with completely observed outcomes, we have $\delta_h = 1$ and $\lambda_{h1}^{(r)} = 1$.
3. Maximize $\Omega\left(\boldsymbol{\psi}, \widehat{\boldsymbol{\psi}}^{(r)}\right) = \sum_{h=1}^{H}\sum_{k=1}^{\delta_h} \lambda_{hk}^{(r)} \ln L_h(\boldsymbol{\psi}|\boldsymbol{O}_h, \boldsymbol{u}_{hl})$ with regard to $\boldsymbol{\psi}$ to find $\widehat{\boldsymbol{\psi}}^{(r+1)}$.
4. Repeat 2 and 3 until convergence in the estimates $\widehat{\boldsymbol{\psi}}^{(r)}$ of $\boldsymbol{\psi}$.

Direct calculation of the covariance matrix is difficult as one need to sum over all possible realizations of the missing data for the whole study population, which is not a linear operation for the calculation of the missing information. Our solution is to estimate the covariance using a sampling approach.[7] Specifically, we sample K sets of missing data $\boldsymbol{U}_h$ for each CCG h from the distribution $Pr\left(\boldsymbol{U}_h|\boldsymbol{O}_h, \widehat{\boldsymbol{\psi}}\right)$, where K is a large integer (e.g., 1000) and $\widehat{\boldsymbol{\psi}}$ is the final parameter estimate. Let these samples be denoted by $\widehat{\boldsymbol{u}}_{hk}$, $h = 1, \cdots, H$, $k = 1, \cdots, K$. The covariance matrix, $\widehat{\boldsymbol{V}}$, is given by

$$\widehat{\boldsymbol{V}}^{-1} = -\frac{1}{K}\sum_{k=1}^{K} \frac{d^2}{d\boldsymbol{\psi}^2} \ln L\left(\boldsymbol{\psi}|\boldsymbol{O}, \widehat{\boldsymbol{U}}_{\cdot\boldsymbol{k}}\right) - \left[\frac{1}{K}\sum_{k=1}^{K}\left(\frac{d}{d\boldsymbol{\psi}} \ln L\left(\boldsymbol{\psi}|\boldsymbol{O}, \widehat{\boldsymbol{U}}_{\cdot\boldsymbol{k}}\right)\right)\left(\frac{d}{d\boldsymbol{\psi}} \ln L\left(\boldsymbol{\psi}|\boldsymbol{O}, \widehat{\boldsymbol{U}}_{\cdot\boldsymbol{k}}\right)\right)' - \left(\frac{1}{K}\sum_{k=1}^{K} \frac{d}{d\boldsymbol{\psi}} \ln L\left(\boldsymbol{\psi}|\boldsymbol{O}, \widehat{\boldsymbol{U}}_{\cdot\boldsymbol{k}}\right)\right)\left(\frac{1}{K}\sum_{k=1}^{K} \frac{d}{d\boldsymbol{\psi}} \ln L\left(\boldsymbol{\psi}|\boldsymbol{O}, \widehat{\boldsymbol{U}}_{\cdot\boldsymbol{k}}\right)\right)'\right],$$

Where $\widehat{\boldsymbol{U}}_{\cdot \boldsymbol{k}} = \{\widehat{\boldsymbol{u}}_{hk} : h = 1, \cdots, H\}$, and $\ln L\left(\boldsymbol{\psi} \middle| \boldsymbol{O}, \widehat{\boldsymbol{U}}_{\cdot \boldsymbol{k}}\right) = \sum_{h=1}^{H} \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \widehat{\boldsymbol{u}}_{hk})$.

In our data, the number of asymptomatic infections in each CCG is at most two, which makes the EM algorithm possible. If there are more asymptomatic infections such that enumeration of all possible realizations in each cluster is impossible, one can use the Monte-Carlo EM (MCEM) algorithm.[7]

1.4 The natural history of disease

The natural history of disease is represented by the density function of the incubation period, $\eta(l)$, and the relative infectivity profile function $\phi(l)$. The incubation period has been estimated to have a mean of 4-7 days with a wide range.[8,9] Although there have been unofficially reported incubation periods as long as more than 20 days, these are likely rare events, and most countries adopted a two-week quarantine policy. For these reasons, we assume a maximum incubation period of 14 days and a minimum duration of 1 day. Based on our unpublished parametric Weibull models for the incubation period for contact tracing clusters of cases in China,[10] we generated four possible incubation period settings with means of 4, 5, 6 and 7 days and standard deviations of $2 \cdot 0$, $2 \cdot 5$, $3 \cdot 0$ and $3 \cdot 5$ days (Table S1).

Much is unknown about the infectious period of COVID-19. In a Germany study of 9 patients, Wölfel et al. found the virus could be cultured up to 8 days post symptom onset (symptom onset counted as day 1 in that paper).[11] In addition, among the 9 patients, viral RNA copies in nasal swabs were above detection at day 8 post symptom onset in 7 (77%) patients and at day 14 in 3 (33%) patients. On the other hand, another study by He et al. using 77 transmission pairs within and outside China suggested relative infectivity declines quickly within 7 days after symptom onset.[1] To accommodate the substantial uncertainty, we chose three settings for the relative infectivity profiles, $\phi(l)$, with maximum durations of 13, 16 and 22 days for the infectious period (Table S2). All the settings include 5 days before symptom onset, and the timeline is in reference to symptom onset as day 0, i.e., $D_{min} = -5$ and D_{max} =7, 10 and 16. As an alternative interpretation, the relative infectivity levels can be thought of as the upper-tail cumulative distribution function for a random interval with constant infectivity, where these settings imply mean infectious periods of 10.1, 12.9 and 16.4 days. The shortest setting hinges with the study by He et al,[1] whereas the longest setting is in line with the German study with relative infectivity of 0.8 at day 7 and 0.3 at day 13 (equivalent to days 8 and 14 if counting onset as day 1).[11] These

relative infectivity levels largely reflect temporal changes in biological infectiousness of a host as immunity develops. Longer infectious periods will not make much difference as most close contacts were isolated by 14 days since symptom onset of primary cases (see Table S3 and Section 1.5 for interpretation of the table). As most estimates of the mean incubation period fall in the range of 5-6 and the He et al. study was based on more subjects than the Wölfel et al. study, we use the setting of a mean incubation period of 5 days and a maximum infectious period of 13 days (with 5 days pre and 7 days post symptom onset) as the primary setting for presenting our findings.

It was suspected that an infected person is actually infectious during the incubation period. This is partially supported by literature as well as our own unpublished findings that the mean serial interval and generation interval are shorter than the mean incubation period.[1,10,12] To account for this possibility, we assume an infected individual can be infectious as early as 5 days before symptom onset (symptomatic case) or peak infectivity (asymptomatic infection). This assumption, together with the assumed incubation period of 1-14 days, implies a latent period of 0-9 days; or more specifically, the latent period is 0 if the incubation period ≤5 days and (incubation period duration – 5 days) otherwise. Although the relative infectivity level may vary during the incubation period, we set it to 1.0 for all pre-onset days for simplicity, i.e., $\phi(l) = 1$ for $l = -5, -4, \cdots, -1$ (i.e., 5, 4 and up to 1 day before symptom onset), as shown in Table S2. For each symptomatic case, we introduce a time-dependent binary indicator takes value 0 for the incubation period and 1 for the illness period starting from the onset day. This illness indicator variable is adjusted for as a covariate affecting infectivity in the regression (1), and its coefficient reflects the difference in infectivity between the incubation period and the illness period. For asymptomatic infections, the illness indicator is 0 throughout the infectious period as there is no illness.

The relative infectivity level during the incubation period of each infection is further adjusted for the uncertainty in the length of the incubation period. The rationale is simple. While we assume an infection can be infectious as early as 5 days before symptom onset or peak infectivity, the incubation period can be shorter than 5 days. Consequently, the effective relative infectivity on day l during the incubation period ($-5 \leq l \leq -1$) is the product of the nominal relative infectivity level $\phi(l)$ and the probability that the incubation period is longer than or equal to $|l|$

days (i.e., the person is infected on or before day l), which is $\phi(l)\sum_{j=|l|}^{d_{max}}\eta(j)$. This adjustment applies to both symptomatic and asymptomatic infections.

1.5 SAR and local reproductive number

We report SAR and local reproductive number estimates for the model adjusting for no covariates except for the time-dependent illness period indicator which affects infectivity of an infected person. For each day l, let $\phi^*(l) = \phi(l)\left[\sum_{j=|l|}^{d_{max}}\eta(j)\right]^{I_{\{l<0\}}}$ be the effective relative infectivity function, and let $p_k^*(l) = (p_k)^{I_{\{l<0\}}}\left(\frac{p_k OR}{1-p_k+p_k OR}\right)^{I_{\{l\geq 0\}}}$ be the effective transmission probability. SAR is calculated as

$$SAR_k = 1 - \prod_{l=D_{min}}^{D_{max}}[1 - p_k^*(l)\phi^*(l)],\ k = 1,2,$$

where OR is the odds ratio for the illness period indicator. The covariate-adjusted transmission probability $\frac{p_k OR}{1-p_k+p_k OR}$ is derived from the logistic regression $\text{logit}^{-1}[\text{logit}(p_k) + \log(OR)]$, where $\text{logit}(p_k) = \log[p/(1-p)]$ and logit^{-1} is the inverse logit transformation. With the illness indicator, p_k is interpreted as the average daily transmission probability from an infective person to a susceptible person ($k = 1$ for household and 2 for non-household) during the incubation period of the infective, and $\frac{p_k OR}{1-p_k+p_k OR}$ is the daily transmission probability when the infective is ill and completely infectious ($\phi(l) = 1$). The SAR among non-household contacts, SAR_2, may be a less appropriate measure for transmissibility than the daily transmission probabilities p_2 or $\frac{p_2 OR}{1-p_2+p_2 OR}$, as some types of non-household contact do not last over the whole infectious period by nature and could even be one-time event, e.g., contact with other restaurant customers or flight passengers. For this reason, we present both p_k and $\frac{p_k OR}{1-p_k+p_k OR}$ in Table S7.

The local reproductive number is defined as the mean number of infections a case can generate during his or her entire infectious period via both close household and non-household contact. This reproductive number can be viewed as an approximate to the basic reproductive number R_0 if the whole population is susceptible and no intervention is implemented. Due to the tight control of human movement in Guangzhou during the study period, this local reproductive number does not reflect R_0 in our study. The distributions of daily numbers of household and

non-household contacts of cases are shown in figure S2, where household is defined by close relatives. Let $n_1(l)$ and $n_2(l)$ be the average numbers of household and non-household contacts per primary case on day l, respectively, with the symptom onset day of the primary case as day 0. The observed values of $n_1(l)$ and $n_2(l)$ are given in Table S3 (2 top rows).

We stop at day 16 as it is the maximum value of D_{max} in our settings for the infectious period. The local reproductive number is calculated as

$$R = \sum_{k=1}^{2} \sum_{l=D_{min}}^{D_{max}} \{n_k(l) p_k^*(l) \phi^*(l) \prod_{m=D_{min}}^{l-1} [1 - p_k^*(l)\phi^*(l)]\},$$

where $n_k(l) \prod_{m=D_{min}}^{l-1} [1 - p_k^*(l)\phi^*(l)]$ is the expected number of susceptible contacts by day $l-1$ among the $n_k(l)$ contacts made on day l. As the observed $n_k(l)$'s are used, this R represents the local reproductive number under the implemented control measures. In particular, the $n_k(l)$'s decreased quickly after day 0 (symptom onset day of the primary case), indicating the effect of isolation or quarantine of both identified cases and their close contacts. To estimate local reproductive numbers in the absence of quarantine (movement constraint may remain) had there been no quarantine, we recalculate R by projecting the average contact numbers during days -5 to 0 to the subsequent days 1 to 16 (Table S3, 2 bottom rows).

1.6 Covariates adjusted in the transmission model

The covariate vector $\boldsymbol{x}_{ij}(t)$ in equation (1) contains characteristics from a susceptible individual i and his or her infectious contact j on day t. Covariates of the susceptible person are considered as affecting susceptibility, and covariates of the infectious person are considered as affecting infectivity. If a covariate is specific to a pair of susceptible and infectious individuals, e.g., household size, then one can treat it as a covariate of either the susceptible person or the infectious person, and the choice is often arbitrary. The following covariates are considered as potential modifiers of either susceptibility or infectivity in the chain-binomial model:

- Age group: 0-19 years old, 20-59 years old, and ≥60 years old (reference). We assume age group affected both infectivity and susceptibility.
- Gender: male and female (reference). We assume gender affected both infectivity and susceptibility.
- Binary indicator for the illness period (1) vs. the incubation period (0).

- Binary indicator for epidemic phase (before Feb. 1, 2020 vs. on or after Feb. 1, 2020).
- Household size (≤6 people vs. >6 people) for within household transmission.

The cut points for epidemic phase (Feb. 1, 2020) and household size (6) were determined by screening data-based and model-based SAR estimates among different dichotomization schemes. Household size (≤6 vs. >6) was not as a predictor in the regression of transmission probability but as a subgroup-specific daily transmission probability, i.e., individuals in households of sizes >6 has a daily household transmission probability different from that in households of sizes ≤6. Consequently, there actually 3 person-to-person transmission probabilities, 2 among household contacts (p_1 for households of sizes ≤6 and p_2 for households of sizes >6) and 1 among non-household contacts (p_3).

The covariate effects on susceptibility apply equally to the external infection probability b and the person-to-person transmission probabilities, except that we do not assume the external infection probability differed by phase. This is because time-dependence of b is not identifiable given the current amount of secondary cases and the relative long infectious period.

The final model used to estimate SAR and local reproductive number is not adjusted for the covariates except for the indicator for illness period vs. incubation period. In the final model used to assess covariate effects (odds ratios), we adjusted susceptibility for age group, epidemic phase and household size, and adjusted infectivity for the indicator for illness period vs. incubation period. Gender was found affecting neither susceptibility nor infectivity, and age group did not change infectivity much; hence, these terms were dropped from the final model.

1.7 Assessment of goodness-of-fit

We had developed a goodness-of-fit measure for chain-binomial models, which compares the observed and model-fitted frequencies of symptom onsets among exposed person-days.[6] The calculation of model-fitted, or expected, frequency of symptom onset for each exposed person-day is conditioning on transmission dynamics that have already realized up to the day before. In the presence of asymptomatic infections, this approach is likely inadequate. Instead, we propose to compare observed and expected frequencies of infection, not symptom onset, for each exposed person-day. The expected frequency for a person-day (i, t) is simply $[1 - e_i(t)]Q_i(t-1)$. The observed frequency is obtained by allocating each observed symptom onset to the possible

infection days, i.e., from $\tilde{t}_i - d_{max}$ to $\tilde{t}_i - d_{min}$, using certain weights. More specifically, the weight for a given $l = \tilde{t}_i - t$, where $t \in [\tilde{t}_i - d_{max}, \tilde{t}_i - d_{min}]$, is given by $w_l = \phi(l)[1 - e_i(t)] \prod_{\tau=\tilde{t}_i - d_{max}}^{t-1} e_i(\tau)$, and the weights are then normalized to have sum 1. These weights for observed frequencies depend on model parameter estimates via $e_i(t)$ in order to account for the fact that the number of infectious individuals on each day should contribute to the allocation of a symptomatic case to the potential infection days of this case. For asymptomatic infections, we choose the most likely peak infectivity day $\tilde{t}_i$ corresponding to value of $\boldsymbol{u}_{hl}$ associated with the highest likelihood at the cluster level, and the calculation of weights proceeds as for symptomatic infections. The observed and expected frequencies of infections are then aggregated over the whole study population by day, where all clusters are aligned with $\tilde{t}_{\Lambda_h} - d_{max} + 1$ as day 1 for each cluster. The 95% confidence intervals (CI) may be calculated for the daily aggregated expected frequencies, but such 95% CIs tend to be overly narrow due to the reduced uncertainty as a result of conditioning the calculation for each exposed person-day on observed transmission dynamics before that day. We adopted a frequently used alternative for constructing marginal, not conditional, 95% CIs by simulating transmissions within clusters using the model-fitted parameters. We plot the observed and expected daily frequencies of infections together with the pointwise 2.5% and 97.5% quantiles of daily numbers of infections in the simulated outbreaks, as a visual diagnosis for the goodness-of-fit of our models.

1.8 Additional sensitivity analyses

We conducted a few additional sensitivity analyses. We first set constant rather than decaying relative infectivity levels over the illness period but restricting the maximum infectious period to be 13 days (Table S11). The SAR estimates slightly increased in comparison to the primary estimates in Table 2 in the main text, and the difference in the relative infectivity between the incubation and illness periods became more dramatic compared to that in Table 3. We then changed the imputation range in the E-M algorithm for the peak infectivity day $\tilde{t}_i$ of each asymptomatic infection i to $(t_i^* - D_{max},\ t_i^* - D_{min})$, where $D_{min} = -5$, D_{max} =7 or 16, and t_i^* is the collection date of the first test-positive nasal swab for individual i. The underlying rational is that t_i^* could be either the first day or the last day of the infectious period $(\tilde{t}_i + D_{min},\ \tilde{t}_i + D_{max})$. Finally, we changed the definition of local primary cases to only local cases with symptom onsets exactly on the earliest symptom onset date in each case cluster. Neither the

change of the imputation range nor the change of the definition of local primary cases affected the SAR estimates much (Table S12), as compared to those in Table 2.

Data and Code Availability Request of sharing deidentified data may be directed to (jingqinlong@126.com), subject to IRB approval at Guangzhou CDC. The R code and data for summary analyses can be downloaded at https://uflorida-my.sharepoint.com/:f:/g/personal/yangyang_ufl_edu/EmJwby1Uj_1LrO9USZCPWWoB6XBQ3lJHoCLGmBZzm-2nEw?e=x8Cuxd. The C code implementing the discrete-time chain-binomial model can be downloaded at https://github.com/yangyang-uf/TranStat.

References

1. He X, Lau EHY, Wu P, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature Med. (2020). https://doi.org/10.1038/s41591-020-0869-5
2. Anderson RM, May RM. Infectious diseases of humans: dynamics and control. Oxford: Oxford University Press; 1991.
3. Nishiura H and Chowell G. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: Chowell G, Hyman JM, Bettencourt LMA, Castillo-Chavez C, editors. Mathematical and statistical estimation approaches in epidemiology. Dordrecht (the Netherlands): Springer Netherlands; 2009. p. 103–21.
4. Wallinga J and Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings of the Royal Society B. 2020; 274:599-604.
5. Farrington CP and Whitaker HJ. Estimation of effective reproduction numbers for infectious diseases using serological survey data. Biostatistics. 2003;4:621–632
6. Yang Y, Longini, IM and Halloran, ME. Design and Evaluation of Prophylactic Intervention Using Infectious Disease Incidence Data from Close Contact Groups. Journal Of the Royal Statistical Society, Series C. 2006; 55: 317-330.
7. Yang Y, Longini IM, Halloran ME and Obenchain V. A hybrid EM and Monte Carlo EM Algorithm and Its Application to Analysis of Transmission of Infectious Diseases. *Biometrics*. 2012; 68: 1238-1249.
8. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. N Engl J Med. 2020;382:1199-1207.
9. Lauer SA, Grantz KH, Bi Q, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020; doi: https://doi.org/10.7326/M20-0504.
10. Lu Q, Zhang Y, Liu M et al. Natural history of disease of the novel coronavirus and its implication for infectivity among patients in China. Under review.
11. Wölfel R, Corman VM, Guggemos W, et al. Virological assessment of hospitalized patients with COVID-2019. Nature (2020). https://doi.org/10.1038/s41586-020-2196-x.
12. Ganyani T, Kremer C, Chen D, Torneri A, Faes C, Wallinga J, et al. Estimating the generation interval for COVID-19 based on symptom onset data. medRxiv 2020:2020.03.05.20031815.

Table S1. Probability densities for the distribution of the incubation period.

Mean Duration (days)	Days from infection to symptom onset													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
4	0·091	0·16	0·19	0·18	0·15	0·10	0·061	0·032	0·015	0·0061	0·0022	$7{\cdot}2\times10^{-4}$	$2{\cdot}1\times10^{-4}$	$5{\cdot}3\times10^{-5}$
5	0·058	0·11	0·14	0·16	0·15	0·13	0·10	0·068	0·044	0·026	0·014	0·0072	0·0034	0·0015
6	0·043	0·079	0·11	0·12	0·13	0·12	0·11	0·088	0·070	0·052	0·037	0·025	0·016	0·0098
7	0·040	0·065	0·082	0·093	0·098	0·098	0·095	0·088	0·080	0·071	0·061	0·052	0·043	0·035

Table S2. Relative infectivity levels during the infectious period. Day 0 corresponds to the symptom onset day for a symptomatic case or the peak infectivity day for an asymptomatic infection.

Max Duration (Days)	Days from symptom onset or peak infectivity day														
	-5 ~ 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
13	1·0	0·8	0·6	0·4	0·2	0·1									
16	1·0	1·0	1·0	0·8	0·8	0·6	0·4	0·2	0·1						
22	1·0	1·0	1·0	1·0	1.0	0·8	0·8	0·6	0·6	0·4	0·4	0·3	0.3	0.1	0.1

Table S3. The average daily numbers of household ($n_1(l)$) and non-household ($n_2(l)$) contacts per primary case on day l of the potential infectious period, with the symptom onset day of the primary case as day 0. Projected numbers from day 1 to day 22 are simply the average of the observed numbers from day -5 to day 0.

	Contact Type	Day during the infectious period																				
		-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Observed	$n_1(l)$	3·57	3·56	3·48	3·47	3·42	3·24	2·68	2·29	1·98	1·45	1·09	0·81	0·63	0·46	0·35	0·2	0·19	0·11	0·1	0·06	0·04	0·01
	$n_2(l)$	2·28	2·24	2·25	2·17	1·9	2·3	1·07	0·92	0·79	0·48	0·31	0·26	0·22	0·16	0·16	0·15	0·04	0·05	0·015	0·01	0·07	0·02
Projected	$n_1(l)$	3·57	3·56	3·48	3·47	3·42	3·24	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46	3·46
	$n_2(l)$	2·28	2·24	2·25	2·17	1·9	2·3	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19	2·19

Table S4. Median and inter-quartile range (IQR) for the days from symptom onset to hospitalization and from symptom onset to laboratory confirmation. The p-values were based on Kruskal-Wallis test for age group and Mann-Whitney U-test for other factors.

Factor	Category	Onset to hospitalization (days)				Onset to confirmation (days)			
		n	Median	IQR	p-value	n	Median	IQR	p-value
Age group	<20 y	19	0	(0, 1)	0.002	20	4	(2, 6)	0.17
	20-59 y	207	2	(0, 5)		224	5	(2, 8)	
	≥60 y	92	2	(0, 5)		95	6	(3, 8)	
Sex	Female	166	2	(0, 5)	0.55	174	5	(3, 8)	0.66
	Male	152	2	(0, 4)		165	5	(3, 8)	
Month	Jan.	263	2	(1, 5)	<0.001	276	6	(4, 9)	<0.001
	Feb.	55	0	(0, 2)		63	3	(1.5, 4)	
Origin	Imported	205	2	(0, 4)	0.16	214	5	(3, 8)	0.5
	Local	113	2	(0, 5)		125	5	(3, 8)	
Case generation	Primary	204	3	(1, 6)	<0.001	209	6	(4, 9)	<0.001
	Secondary	114	1	(0, 2)		130	4	(2, 6)	
Total		318	2	(0, 5)		339	5	(3, 8)	

Table S5. Frequency (percentage) of key symptoms of COVID-19 cases identified by case-finding and contact-tracing in Guangzhou, China up to Feb. 18, 2020.

Factor	Category	n	Systematic Symptoms					Respiratory Symptoms					Chest CT/X-ray
			Fever	Fatigue	Chills	Myalgia	Head-ache	Cough	Sore Throat	Runny Nose	Short Breath	Diarrhea	
Age group	<20 y	20	13 (65)	1 (5)	2 (10)	0 (0)	2 (10)	10 (50)	4 (20)	6 (30)	0 (0)	1 (5)	8 (62)
	20-59 y	234	170 (73)	47 (20)	35 (15)	41 (18)	35 (15)	138 (59)	45 (19)	26 (11)	20 (9)	19 (8)	160 (79)
	≥60 y	95	75 (79)	26 (27)	18 (19)	10 (11)	9 (9)	59 (62)	11 (12)	4 (4)	10 (11)	5 (5)	70 (86)
Sex	Female	181	129 (71)	33 (18)	28 (15)	25 (14)	24 (13)	105 (58)	37 (20)	17 (9)	20 (11)	16 (9)	120 (80)
	Male	168	129 (77)	41 (24)	27 (16)	26 (15)	22 (13)	102 (61)	23 (14)	19 (11)	10 (6)	9 (5)	118 (81)
Origin	Imported	220	168 (76)	40 (18)	39 (18)	16 (12)	25 (11)	133 (60)	42 (19)	18 (8)	13 (6)	14 (6)	155 (80)
	Local	129	90 (70)	34 (26)	16 (12)	25 (19)	21 (16)	74 (57)	18 (14)	18 (14)	17 (13)	11 (9)	83 (81)
Case Type	Primary	215	178 (83)	53 (25)	48 (22)	36 (17)	35 (16)	136 (63)	39 (18)	26 (12)	20 (9)	21 (10)	169 (86)
	Secondary	134	80 (60)	21 (16)	7 (5)	15 (11)	11 (8)	71 (53)	21 (16)	10 (7)	10 (7)	4 (3)	69 (70)
Total		349	258 (74)	74 (21)	55 (16)	51 (15)	46 (13)	207 (59)	60 (17)	36 (10)	30 (9)	25 (7)	238 (80)

Table S6. Model-based estimates (and 95% confidence intervals) of secondary attack rates (SAR) among household and non-household contacts, and model-based estimates of local reproductive number (local R) with and without quarantine. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group, epidemic phase or household size.

Mean Incubation Period (days)	Max Infectious Period (days)	Household defined by close relatives				Household defined by residential address			
		SAR （%）		Local Reproductive Number		SAR （%）		Local Reproductive Number	
		Household	Non-household	With quarantine	Without quarantine	Household	Non-household	With quarantine	Without quarantine
4	13	13·3 (10·6-16·5)	8·4 (5·6-12·4)	0·49 (0·4-0·6)	0·64 (0·52-0·79)	18·2 (14·2-22·9)	7·9 (5·8-10·6)	0·49 (0·4-0·6)	0·63 (0·51-0·77)
	16	15·2 (12·0-19·1)	9·9 (6·6-14·7)	0·5 (0·4-0·62)	0·74 (0·59-0·92)	20·7 (16·0-26·4)	9·1 (6·6-12·4)	0·5 (0·4-0·62)	0·72 (0·58-0·9)
	22	18·0 (13·9-23·0)	12·2 (8·0-18·1)	0·5 (0·4-0·64)	0·89 (0·7-1·13)	24·3 (18·5-31·2)	11·0 (7·8-15·2)	0·5 (0·39-0·63)	0·86 (0·67-1·09)
5	13	12·4 (9·8-15·4)	7·9 (5·3-11·8)	0·5 (0·41-0·62)	0·6 (0·49-0·74)	17·1 (13·3-21·8)	7·3 (5·4-9·9)	0·5 (0·4-0·61)	0·59 (0·48-0·72)
	16	13·6 (10·6-17·3)	8·9 (5·9-13·4)	0·5 (0·4-0·64)	0·67 (0·53-0·84)	18·8 (14·4-24·2)	8·1 (5·8-11·1)	0·5 (0·4-0·63)	0·65 (0·52-0·82)
	22	15·5 (11·7-20·2)	10·4 (6·7-15·8)	0·51 (0·39-0·66)	0·76 (0·59-1)	21·2 (15·8-27·8)	9·3 (6·5-13·1)	0·5 (0·38-0·65)	0·74 (0·57-0·96)
6	13	11·7 (9·3-14·6)	7·7 (5·1-11·3)	0·5 (0·41-0·62)	0·57 (0·46-0·71)	16·4 (12·7-20·8)	7·0 (5·1-9·4)	0·5 (0·41-0·62)	0·56 (0·46-0·69)
	16	12·6 (9·8-16·0)	8·4 (5·5-12·6)	0·51 (0·4-0·64)	0·62 (0·49-0·78)	17·6 (13·4-22·6)	7·5 (5·4-10·3)	0·5 (0·4-0·63)	0·6 (0·48-0·76)
	22	13·8 (10·4-18·2)	9·3 (6·0-14·3)	0·51 (0·39-0·67)	0·68 (0·52-0·9)	19·2 (14·2-25·3)	8·3 (5·8-11·7)	0·5 (0·38-0·66)	0·66 (0·51-0·86)
7	13	11·4 (9·0-14·2)	7·5 (5·0-11·2)	0·51 (0·41-0·63)	0·56 (0·45-0·69)	16·1 (12·5-20·4)	6·8 (5·0-9·2)	0·5 (0·41-0·62)	0·55 (0·45-0·67)
	16	12·1 (9·5-15·3)	8·1 (5·3-12·2)	0·51 (0·41-0·65)	0·6 (0·47-0·75)	17 (13·1-21·9)	7·2 (5·2-9·9)	0·51 (0·4-0·63)	0·58 (0·47-0·73)
	22	13·1 (9·9-17·1)	8·9 (5·7-13·6)	0·51 (0·39-0·67)	0·65 (0·49-0·85)	18·3 (13·6-24·1)	7·8 (5·5-11)	0·51 (0·39-0·66)	0·63 (0·48-0·82)

Table S7. Model-based estimates (and 95% confidence intervals) of daily transmission probabilities for household contacts (p_1^*) and non-household contacts (p_2^*) during the incubation and illness periods. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group, epidemic phase or household size.

Mean Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives				Household contact defined by residential address			
		Incubation Period		Illness Period		Incubation Period		Illness Period	
		$p_1^* \times 10^{-2}$	$p_2^* \times 10^{-2}$	$p_1^* \times 10^{-2}$	$p_2^* \times 10^{-2}$	$p_1^* \times 10^{-2}$	$p_2^* \times 10^{-2}$	$p_1^* \times 10^{-2}$	$p_2^* \times 10^{-2}$
4	13	1·6 (1·12-2·28)	0·99 (0·6-1·62)	1·66 (1·07-2·56)	1·03 (0·59-1·77)	2·27 (1·56-3·31)	0·94 (0·62-1·42)	2·3 (1·46-3·6)	0·95 (0·58-1·55)
	16	1·68 (1·2-2·34)	1·07 (0·66-1·71)	1·31 (0·84-2·04)	0·83 (0·48-1·45)	2·4 (1·68-3·43)	1 (0·67-1·47)	1·83 (1·16-2·88)	0·75 (0·46-1·24)
	22	1·7 (1·22-2·35)	1·11 (0·7-1·76)	1·2 (0·78-1·85)	0·79 (0·45-1·36)	2·43 (1·71-3·45)	1·02 (0·7-1·5)	1·66 (1·05-2·61)	0·7 (0·42-1·14)
5	13	1·84 (1·36-2·49)	1·16 (0·73-1·83)	1·13 (0·61-2·08)	0·71 (0·35-1·43)	2·64 (1·9-3·66)	1·08 (0·75-1·55)	1·58 (0·84-2·95)	0·64 (0·33-1·24)
	16	1·9 (1·42-2·53)	1·22 (0·78-1·89)	0·89 (0·48-1·62)	0·57 (0·28-1·14)	2·74 (2-3·74)	1·12 (0·79-1·58)	1·24 (0·66-2·29)	0·5 (0·26-0·96)
	22	1·91 (1·44-2·54)	1·25 (0·81-1·92)	0·8 (0·44-1·46)	0·52 (0·26-1·05)	2·77 (2·03-3·76)	1·14 (0·81-1·61)	1·1 (0·59-2·05)	0·45 (0·23-0·87)
6	13	2·01 (1·53-2·64)	1·3 (0·84-2)	0·74 (0·31-1·75)	0·47 (0·18-1·21)	2·9 (2·15-3·9)	1·18 (0·84-1·65)	1·05 (0·44-2·49)	0·42 (0·17-1·03)
	16	2·04 (1·57-2·64)	1·33 (0·87-2·02)	0·59 (0·26-1·33)	0·38 (0·16-0·94)	2·95 (2·22-3·92)	1·2 (0·87-1·66)	0·83 (0·37-1·88)	0·33 (0·14-0·78)
	22	2·05 (1·59-2·64)	1·35 (0·89-2·04)	0·53 (0·24-1·18)	0·35 (0·14-0·85)	2·98 (2·24-3·94)	1·22 (0·89-1·67)	0·74 (0·33-1·67)	0·3 (0·13-0·7)
7	13	2·09 (1·63-2·69)	1·37 (0·9-2·07)	0·54 (0·19-1·57)	0·35 (0·11-1·09)	3·03 (2·29-4)	1·23 (0·89-1·69)	0·79 (0·28-2·21)	0·32 (0·11-0·91)
	16	2·1 (1·65-2·68)	1·39 (0·92-2·08)	0·45 (0·17-1·18)	0·3 (0·1-0·83)	3·06 (2·33-4·01)	1·24 (0·91-1·69)	0·64 (0·25-1·65)	0·26 (0·1-0·68)
	22	2·11 (1·66-2·68)	1·4 (0·93-2·1)	0·41 (0·16-1·05)	0·27 (0·1-0·75)	3·07 (2·35-4·02)	1·25 (0·92-1·7)	0·57 (0·22-1·46)	0·23 (0·09-0·61)

Table S8. Model-based estimates (and 95% confidence intervals) of daily probability of infection from an external source (b) and the odds ratios for the relative infectivity during the illness versus incubation period. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group, epidemic phase or household size.

Mean Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives		Household contact defined by residential address	
		b ($\times 10^{-4}$)	Odds Ratio	b ($\times 10^{-4}$)	Odds Ratio
4	13	2·02 (1·00, 4·07)	1·04 (0·52, 2·06)	2·04 (1·02, 4·09)	1·01 (0·51, 2·01)
	16	1·82 (0·86, 3·85)	0·78 (0·40, 1·51)	1·83 (0·87, 3·85)	0·76 (0·39, 1·47)
	22	1·72 (0·80, 3·67)	0·70 (0·37, 1·35)	1·75 (0·82, 3·71)	0·68 (0·35, 1·30)
5	13	1·71 (0·78, 3·78)	0·61 (0·27, 1·38)	1·74 (0·79, 3·84)	0·59 (0·26, 1·35)
	16	1·54 (0·67, 3·54)	0·46 (0·21, 1·01)	1·58 (0·69, 3·61)	0·44 (0·20, 0·98)
	22	1·49 (0·65, 3·44)	0·41 (0·19, 0·89)	1·54 (0·67, 3·53)	0·39 (0·18, 0·86)
6	13	1·55 (0·64, 3·72)	0·36 (0·13, 1·02)	1·55 (0·64, 3·77)	0·36 (0·13, 1·00)
	16	1·41 (0·57, 3·50)	0·29 (0·11, 0·75)	1·43 (0·57, 3·55)	0·28 (0·11, 0·72)
	22	1·38 (0·56, 3·42)	0·26 (0·10, 0·66)	1·41 (0·57, 3·50)	0·24 (0·09, 0·63)
7	13	1·54 (0·61, 3·86)	0·26 (0·08, 0·86)	1·53 (0·60, 3·87)	0·26 (0·08, 0·82)
	16	1·41 (0·55, 3·63)	0·21 (0·07, 0·63)	1·42 (0·55, 3·66)	0·20 (0·07, 0·59)
	22	1·38 (0·54, 3·56)	0·19 (0·07, 0·55)	1·40 (0·54, 3·62)	0·18 (0·06, 0·52)

Table S9. Model-based odds ratios (and 95% confidence intervals) for effects of age group and epidemic phase on susceptibility and infectivity and relative infectivity during the illness period in comparison to the incubation period. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is adjusted for age group, epidemic phase and household size

Definition of Household	Mean Incubation Period (days)	Max Infectious Period (days)	Covariates Affecting			
			Susceptibility			Infectivity
			<20 vs. ≥60	20-59 vs. ≥60	Feb. vs. Jan.	Illness vs. Incubation
Close Relatives	4	13	0·23 (0·11-0·47)	0·65 (0·44-0·98)	0·50 (0·22-1·16)	0·98 (0·5-1·94)
		16	0·23 (0·11-0·47)	0·65 (0·43-0·97)	0·53 (0·24-1·18)	0·75 (0·39-1·46)
		22	0·23 (0·11-0·47)	0·64 (0·43-0·96)	0·52 (0·24-1·13)	0·69 (0·36-1·33)
	5	13	0·23 (0·11-0·46)	0·64 (0·43-0·97)	0·42 (0·17-1·07)	0·60 (0·27-1·36)
		16	0·22 (0·11-0·46)	0·64 (0·43-0·96)	0·45 (0·19-1·12)	0·46 (0·21-1·01)
		22	0·22 (0·11-0·46)	0·64 (0·42-0·96)	0·46 (0·19-1·1)	0·42 (0·19-0·91)
	6	13	0·22 (0·11-0·46)	0·64 (0·42-0·95)	0·37 (0·14-1·03)	0·39 (0·14-1·04)
		16	0·22 (0·11-0·45)	0·63 (0·42-0·95)	0·40 (0·15-1·07)	0·30 (0·12-0·77)
		22	0·22 (0·11-0·45)	0·63 (0·42-0·95)	0·41 (0·15-1·08)	0·27 (0·11-0·68)
	7	13	0·22 (0·11-0·45)	0·63 (0·42-0·95)	0·36 (0·12-1·05)	0·29 (0·1-0·88)
		16	0·22 (0·11-0·45)	0·63 (0·42-0·95)	0·37 (0·13-1·08)	0·23 (0·08-0·65)
		22	0·22 (0·11-0·45)	0·63 (0·42-0·94)	0·38 (0·13-1·09)	0·21 (0·07-0·58)
Residential Address	4	13	0·23 (0·11-0·47)	0·68 (0·46-1·02)	0·67 (0·3-1·51)	0·93 (0·46-1·86)
		16	0·23 (0·11-0·47)	0·68 (0·46-1·01)	0·71 (0·33-1·55)	0·70 (0·36-1·38)
		22	0·23 (0·11-0·46)	0·68 (0·45-1·01)	0·70 (0·33-1·5)	0·65 (0·34-1·26)
	5	13	0·22 (0·11-0·46)	0·67 (0·45-1)	0·57 (0·23-1·39)	0·54 (0·23-1·26)
		16	0·22 (0·11-0·45)	0·67 (0·45-1)	0·62 (0·26-1·45)	0·42 (0·19-0·94)
		22	0·22 (0·11-0·45)	0·67 (0·45-1)	0·62 (0·27-1·44)	0·38 (0·17-0·84)
	6	13	0·22 (0·11-0·45)	0·67 (0·45-1)	0·51 (0·2-1·33)	0·33 (0·11-0·96)
		16	0·22 (0·11-0·44)	0·67 (0·45-1)	0·55 (0·22-1·39)	0·27 (0·1-0·7)
		22	0·22 (0·11-0·44)	0·67 (0·45-1)	0·56 (0·22-1·41)	0·24 (0·09-0·63)
	7	13	0·22 (0·11-0·44)	0·66 (0·44-0·99)	0·50 (0·18-1·36)	0·24 (0·07-0·79)
		16	0·22 (0·11-0·44)	0·66 (0·44-0·99)	0·52 (0·19-1·4)	0·20 (0·07-0·59)
		22	0·22 (0·11-0·44)	0·66 (0·44-0·99)	0·53 (0·2-1·42)	0·18 (0·06-0·52)

Table S10. Model-based estimates (and 95% confidence intervals) of daily probability of infection from an external source (b) and daily transmission probabilities during the incubation period of a COVID-19 case for household contacts in households of sizes ≤6 (p_1) and in households of sizes >6 (p_2), and for non-household contacts (p_3). Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is adjusted for age group, epidemic phase and household size.

Definition of Household	Mean Incubation Period (days)	Max Infectious Period (days)	b ($\times 10^{-4}$)	p_1 ($\times 10^{-2}$)	p_2 ($\times 10^{-2}$)	p_3 ($\times 10^{-2}$)
Close Relatives	4	13	2·90 (1·36-6·2)	3·65 (2·24-5·89)	1·82 (1·09-3·02)	1·77 (0·96-3·21)
		16	2·62 (1·17-5·87)	3·83 (2·38-6·09)	1·91 (1·16-3·12)	1·87 (1·04-3·35)
		22	2·51 (1·11-5·67)	3·87 (2·42-6·13)	1·94 (1·18-3·15)	1·96 (1·1-3·46)
	5	13	2·45 (1·04-5·77)	4·21 (2·68-6·57)	2·06 (1·27-3·32)	2·11 (1·19-3·71)
		16	2·22 (0·91-5·41)	4·35 (2·8-6·7)	2·13 (1·33-3·38)	2·18 (1·25-3·79)
		22	2·18 (0·89-5·32)	4·39 (2·83-6·75)	2·15 (1·35-3·41)	2·24 (1·29-3·87)
	6	13	2·14 (0·82-5·59)	4·61 (3.0-7·03)	2·22 (1·4-3·5)	2·37 (1·37-4·07)
		16	2.0 (0·75-5·28)	4·70 (3·08-7·11)	2·26 (1·44-3·54)	2·42 (1·41-4·12)
		22	1·98 (0·75-5·22)	4·73 (3·11-7·14)	2·29 (1·46-3·56)	2·45 (1·43-4·17)
	7	13	2·10 (0·78-5·68)	4·82 (3·17-7·27)	2·29 (1·46-3·58)	2·48 (1·45-4·22)
		16	1·98 (0·72-5·4)	4·88 (3·23-7·32)	2·32 (1·49-3·6)	2·52 (1·48-4·26)
		22	1·97 (0·72-5·36)	4·91 (3·25-7·35)	2·34 (1·51-3·62)	2·55 (1·5-4·3)
Residential Address	4	13	2·97 (1·41-6·26)	4·60 (2·82-7·41)	2·12 (1·04-4·29)	1·54 (0·92-2·57)
		16	2·71 (1·23-5·97)	4·79 (2·98-7·61)	2·32 (1·16-4·57)	1·61 (0·98-2·63)
		22	2·55 (1·14-5·69)	4·81 (3·01-7·61)	2·37 (1·2-4·63)	1·65 (1·01-2·67)
	5	13	2·56 (1·11-5·93)	5·35 (3·39-8·33)	2·54 (1·29-4·93)	1·80 (1·12-2·87)
		16	2·31 (0·96-5·58)	5·46 (3·51-8·42)	2·72 (1·41-5·16)	1·83 (1·15-2·9)
		22	2·24 (0·92-5·43)	5·50 (3·54-8·45)	2·75 (1·43-5·2)	1·86 (1·18-2·92)
	6	13	2·28 (0·89-5·81)	5·88 (3·81-8·97)	2·81 (1·47-5·34)	1·97 (1·26-3·07)
		16	2·07 (0·79-5·46)	5·92 (3·87-8·96)	2·95 (1·56-5·5)	1·98 (1·28-3·07)
		22	2·03 (0·77-5·35)	5·95 (3·9-8·98)	2·97 (1·58-5·53)	2.0 (1·29-3·08)
	7	13	2·25 (0·85-5·94)	6·16 (4·04-9·28)	2·94 (1·55-5·51)	2·04 (1·32-3·14)
		16	2·05 (0·75-5·59)	6·16 (4·06-9·23)	3·05 (1·63-5·63)	2·04 (1·33-3·13)
		22	2·02 (0·74-5·49)	6·18 (4·09-9·24)	3·07 (1·65-5·65)	2·05 (1·34-3·14)

Table S11. Sensitivity analysis: model-based estimates (and 95% confidence intervals) of secondary attack rates (SAR) among household and non-household contacts and the odds ratios for the relative infectivity during the illness versus incubation period, under constant rather than time-varying relative infectivity ($\phi(l)$) during the illness period. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group, epidemic phase or household size.

Mean Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives			Household contact defined by residential address		
		Household SAR (%)	Non-household SAR (%)	Odds Ratio	Household SAR (%)	Non-household SAR (%)	Odds Ratio
4	13	14·9 (11·7-18·8)	9·7 (6·4-14·5)	0·73 (0·37, 1·42)	20·4 (15·7-26·0)	9·0 (6·5-12·2)	0·70 (0·36, 1·38)
5	13	13·5 (10·5-17·1)	8·8 (5·8-13·2)	0·44 (0·20, 0·96)	18·6 (14·3-23·9)	8·0 (5·8-11·0)	0·42 (0·19, 0·93)
6	13	12·5 (9·7-15·8)	8·3 (5·4-12·4)	0·27 (0·10, 0·71)	17·4 (13·4-22·4)	7·4 (5·4-10·2)	0·26 (0·10, 0·68)
7	13	12·0 (9·4-15·2)	8·1 (5·3-12·1)	0·20 (0·07, 0·60)	16·9 (13-21·7)	7·2 (5·2-9·8)	0·20 (0·07, 0·57)

Table S12. Sensitivity analysis: model-based estimates (and 95% confidence intervals) of secondary attack rates (SAR) among household and non-household contacts for two sensitivity settings: (1) The imputation range in the E-M algorithm for the peak infectivity day $\tilde{t}_i$ of each asymptomatic infection i was changed to $(t_i^* - D_{max},\ t_i^* - D_{min})$, where $D_{min} = -5$, D_{max} =7 or 16, and t_i^* is the collection date of the first test-positive nasal swab for individual i; (2) The definition of local primary cases was changed to local cases with symptom onsets exactly on the earliest symptom onset date in each case cluster. These models are not adjusted for age group, epidemic phase or household size.

Change	Definition of household	Setting	Mean incubation period = 5 days		Mean incubation period = 7 days	
			Max infectious period = 13 days	Max infectious period = 22 days	Max infectious period = 13 days	Max infectious period = 22 days
Imputation range for asymptomatic infection	Close relatives	Household	12·4 (9·9, 15·5)	15·7 (11·8, 20·5)	11·3 (9·0, 14·1)	12·9 (9·6, 17·1)
		Non-household	7·8 (5·2, 11·6)	10·2 (6·5, 15·8)	7·3 (4·9, 10·9)	8·5 (5·3, 13·3)
	Residential address	Household	17·3 (13·4, 21·9)	21·5 (16·0, 28·3)	16·0 (12·5, 20·4)	18·1 (13·4, 24·1)
		Non-household	7·3 (5·4, 9·9)	9·3 (6·4, 13·1)	6·7 (4·9, 9·0)	7·5 (5·2, 10·8)
Definition of local primary case	Close relatives	Household	12·6 (10·1, 15·6)	15·5 (11·8, 20·2)	11·6 (9·3, 14·4)	13·3 (10·1, 17·2)
		Non-household	7·9 (5·3, 11·6)	10·1 (6·5, 15·3)	7·6 (5·1, 11·1)	8·7 (5·7, 13·3)
	Residential address	Household	17·0 (13·3, 21·6)	20·8 (15·6, 27·2)	16·1 (12·6, 20·3)	18·2 (13·7, 23·7)
		Non-household	7·6 (5·7, 10·2)	9·5 (6·7, 13·3)	7·2 (5·3, 9·5)	8·1 (5·8, 11·3)

Figure S1. Spatial distribution of COVID-19 case clusters with at least one secondary cases. For each cluster, the outer ring indicates whether the primary case was imported (red) or local (blue) and whether the symptom onset of the primary case was on/before (dark color) or after (light color) Jan. 23, 2020, the day when lockdown of Wuhan was initiated. The proportions of household (green) and non-household (purple) cases are shown in the inner circle for each cluster. Township-level population densities are shown as the background.

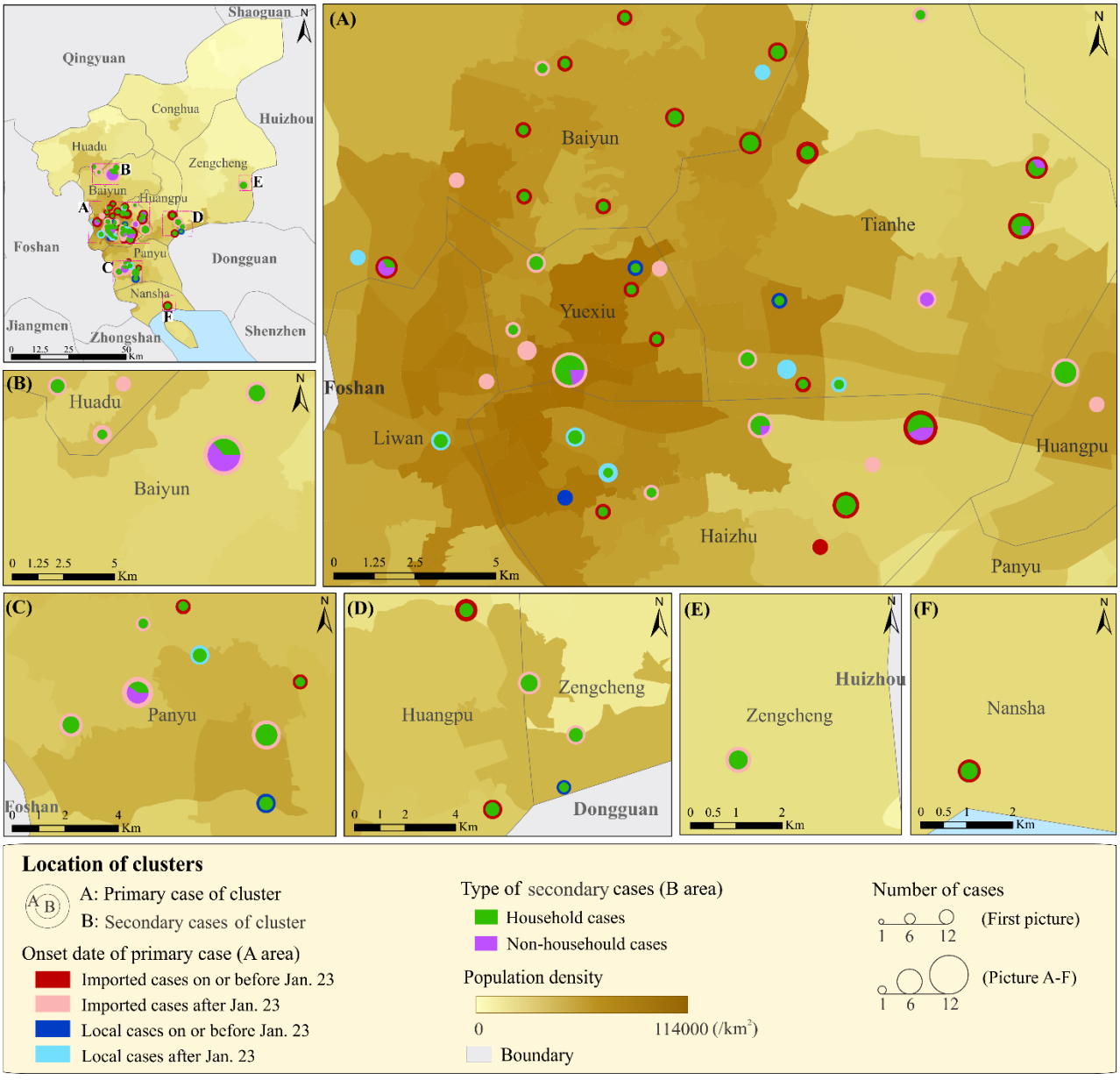

Figure S2. Distribution of number of household (upper) and non-household (lower, log-scale) contacts across all close contact groups in Guangzhou, China over the potential infectious period of the primary case with the symptom onset day of the primary case set as day 0. For close contact groups with more than 1 primary cases, the numbers of contacts are averaged over the co-primary cases. Each box shows the median (middle line) and the inter-quartile range.

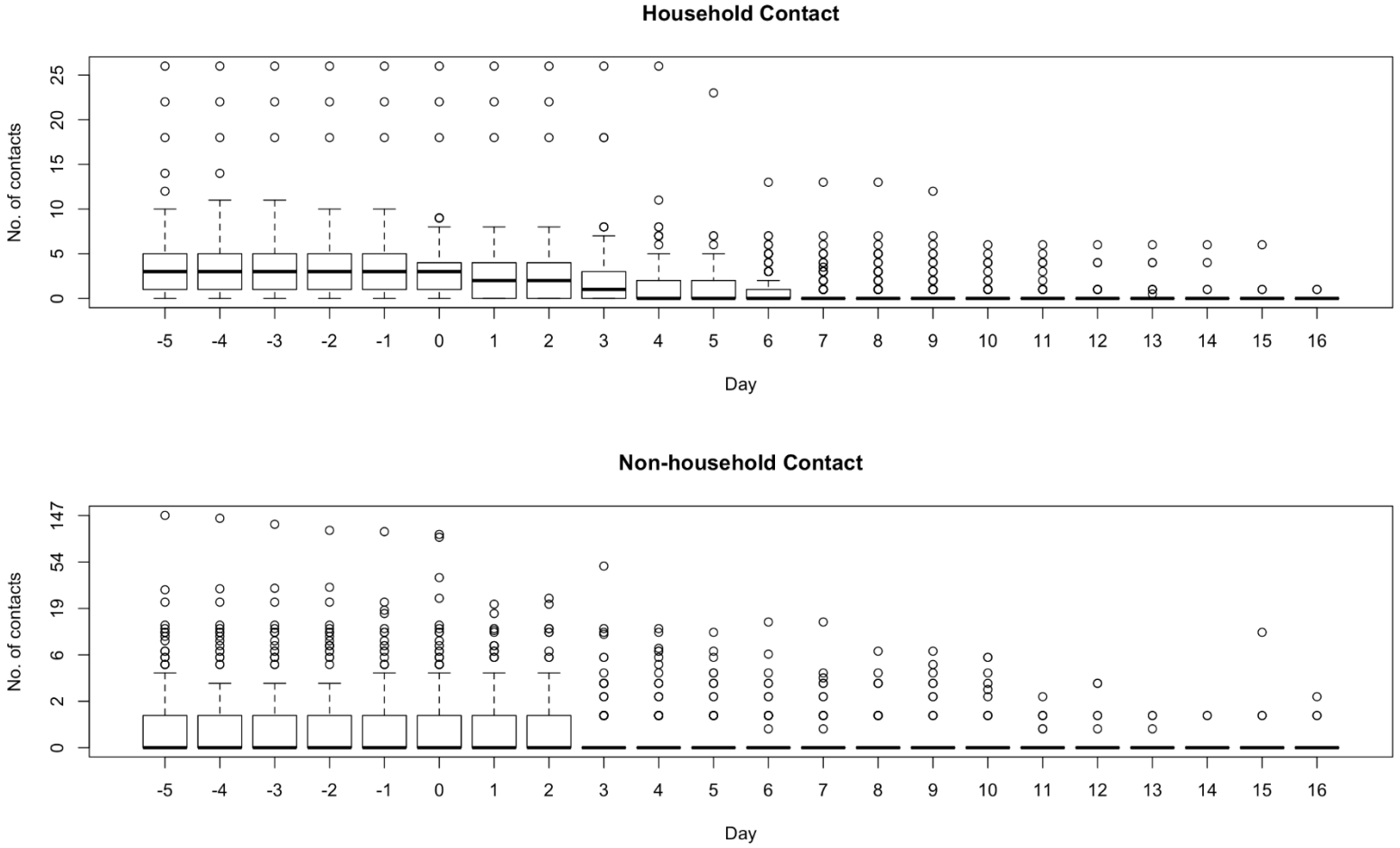

Figure S3. Observed (red triangles) and model-fitted (red solid curve) numbers of secondary infections, together with the 95% CIs for the model-fitted numbers (blue dashed curves), on each day for the whole study population for different settings of the incubation and infectious periods. All close contact groups were aligned in time by the symptom onset day of the earliest primary case and day 0 is 13 days before that day. The observed daily infection numbers changed slightly by the settings of the incubation and infectious periods because the way we allocate cases to their possible infection days also depend on the settings and model parameter estimates (see Appendix 1.7)